Psychology and Aging
2008, Vol. 23, No. 2, 366–376

0882-7974/08/$12.00 DOI: 10.1037/0882-7974.23.2.366

Vocabulary Test Format and Differential Relations to Age

Ryan P. Bowles
Michigan State University

Timothy A. Salthouse
University of Virginia

Although vocabulary tests are generally considered interchangeable, regardless of format, different tests can have different relations to age and to other cognitive abilities. In this study, 4 vocabulary test formats were examined: multiple-choice synonyms, multiple-choice antonyms, produce the definition, and picture identification. Results indicated that, although they form a single coherent vocabulary knowledge factor, the formats have different relations to age. In earlier adulthood, picture identification had the strongest growth, and produce the definition had the weakest. In later adulthood, picture identification had the strongest decline, and multiple-choice synonyms had the least. The formats differed in their relation to other cognitive variables, including reasoning, spatial visualization, memory, and speed. After accounting for the differential relations to other cognitive variables, differences in relation to age were eliminated with the exception of differences for the picture identification test. No theory of the aging of vocabulary knowledge fully explains these findings. These results suggest that using a single indicator of vocabulary may yield incomplete and somewhat misleading results about the aging of vocabulary knowledge.

Keywords: vocabulary knowledge, aging, measurement

Regardless of format, vocabulary tests are generally considered interchangeable indicators of vocabulary knowledge, a principle Spearman (1927) coined *indifference of the indicator*. In his comprehensive analyses of the factor structure of human abilities, Carroll (1993) concluded, "The precise format by which vocabulary knowledge is measured generally makes little difference in the factorial composition of the variables, to the extent that the underlying trait being measured is range of native-language vocabulary knowledge" (p. 158). Intelligence test batteries almost always contain at least one test described as a test of vocabulary, but they can differ markedly in the format used, such as definition production on the Wechsler Adult Intelligence Scale–Third Edition (WAIS-III; Wechsler, 1997a), picture and word identification on the Woodcock–Johnson Psycho-Educational Battery–Revised (WJ-R; Woodcock & Johnson, 1990), and multiple-choice synonyms on the Educational Testing Service Kit (Ekstrom, French, Harman, & Derman, 1976). Some test batteries even include more than one type of vocabulary test, but they are invariably treated as measures of the same vocabulary ability construct (e.g. Munoz-Sandoval, Cummins, Alvarado, & Ruef, 1998; Woodcock, 1987; Woodcock & Johnson, 1990).

Furthermore, the general shape of the curve relating age to vocabulary knowledge seems to be independent of the particular task used. Studies using various formats, including multiple-choice tests (Alwin & McCammon, 2001; Schaie, 1996), identification tests (McGrew & Woodcock, 2001), and production tests (McArdle, Grimm, Hamagami, Bowles, & Meredith, 2008), have indicated that vocabulary knowledge increases throughout early adulthood, flattens out in middle age (ages 40–60), then holds steady or declines gradually in late adulthood (Singer, Verhaeghen, Ghisletta, Lindenberger, & Baltes, 2003). This contrasts sharply with other intellectual abilities, for which the peak ability in cross-sectional studies occurs around age 20 (e.g., Schaie, 1996). The size of the increase in vocabulary ability between early and middle or late adulthood is substantial; in a meta-analysis of 324 studies, Verhaeghen (2003) found that older adults (mean age = 70.4) scored approximately 0.8 standard deviations above younger adults (mean age = 21.4).

Despite the treatment of the relation between age and vocabulary knowledge as independent of the vocabulary test format, some research has indicated that scores from different types of vocabulary knowledge tests have different relations to age (Sorenson, 1938; Verhaeghen, 2003). Although Verhaeghen (2003), in particular, speculated about some theoretical processing differences accounting for certain differential age relations, no research has systematically addressed this issue. Furthermore, studies addressing differences across vocabulary format have examined only two types of vocabulary tests (e.g., synonyms vs. antonyms, Sorenson, 1938; produce the definition vs. multiple choice, Verhaeghen, 2003).

The goal of this study is to systematically examine differences in the relation between age and vocabulary knowledge across four different test formats. For both practical and theoretical reasons, identifying and understanding the differential age relations is important for researchers studying aging:

1. Identifying the sources of differences among vocabulary test formats may inform understanding of the construct of

Ryan P. Bowles, Department of Psychology, Michigan State University; Timothy A. Salthouse, Department of Psychology, University of Virginia.

We acknowledge the help provided by John J. McArdle in analyzing the data. Support for the preparation of this article was provided by Grants T32 AG20500-01 and RO1 AG019627 from the National Institute on Aging.

Correspondence concerning this article should be addressed to Ryan P. Bowles, Department of Psychology, Michigan State University, 298C Psychology Building, East Lansing, MI 48824-1116. E-mail: bowlesr@msu.edu

vocabulary knowledge in general, as well as theories about the aging of vocabulary.

2. Differences among vocabulary test formats may highlight differences in the specific processes required to solve individual vocabulary items and the relation of these processes to age.

3. Most studies incorporating vocabulary knowledge use only a single vocabulary test format. Results based on a single format may be misleading because that format reflects not just what is common about vocabulary knowledge, but also what is unique to the specific test format. Understanding the differences among formats may provide researchers with better tools for interpreting and critiquing results based on a single format.

4. It may be possible to identify the vocabulary test format that most closely matches the average or typical vocabulary test. Researchers may then be more confident that results based on that single indicator are accurate representations of the role of vocabulary knowledge.

In this study, we addressed each of these issues with four types of vocabulary tests: a locally developed multiple-choice synonyms test (Salthouse, 1993), a locally developed multiple-choice antonyms test (Salthouse, 1993), the WAIS-III Vocabulary produce-the-definition test (Wechsler, 1997a), and the WJ-R Picture Vocabulary picture-identification test (Woodcock & Johnson, 1990). We first examined the magnitude of differences in the formats' relations to age. We then attempted to identify sources of the differential age relations by examining relations to other cognitive abilities. The cognitive variables formed the basis of a mediational approach, in which we examined whether the differential relations to other cognitive abilities accounted for the differential age relations. Finally, we offer interpretations of these findings in terms of the four issues described earlier.

Method

Participants

The data were obtained from 3,512 persons who participated in 1 of 18 previously published studies by Salthouse and colleagues in which at least two vocabulary tests were administered (Hambrick, Salthouse, & Meinz, 1999, Studies 1, 2, 3, and 4; Meinz & Salthouse, 1998; Salthouse, 1996, 2001a, Studies 1 and 2; Salthouse, 2001b; Salthouse, Atkinson, & Berish, 2003; Salthouse & Ferrer-Caja, 2003; Salthouse, Fristoe, McGuthry, & Hambrick, 1998, Study 2; Salthouse, Hambrick, Lukas, & Dell, 1996, Study 2; Salthouse, Hancock, Meinz, & Hambrick, 1996, Study 3; Salthouse, McGuthry, & Hambrick, 1999; Salthouse et al., 2000, Study 2; Salthouse, Toth, Hancock, & Woodard, 1997; Siedlecki, Salthouse, & Berish, 2005). Participants ranged in age from 18 to 98 ($M = 49.5$, $SD = 17.2$). Health and education levels of the participants are presented in Table 1.

Procedures

In each study, participants were administered at least two vocabulary tests, as well as a number of other cognitive tasks that varied across studies. A selection of these cognitive tasks was made, with a goal of having several tasks in a number of broad cognitive abilities. A task was selected only if it was used in at least two studies. Eighteen tasks met the selection criteria, resulting in four broad cognitive abilities: reasoning, spatial visualiza-

Table 1
Demographic Characteristics of Participants

Study	*N*	Education	Health
Hambrick, Salthouse, and Meinz (1999), Study 1	202	3.4_a	2.4
Hambrick et al. (1999), Study 2	218	16.1	2.0
Hambrick et al. (1999), Study 3	195	4.0_a	1.9
Hambrick et al. (1999), Study 4	200	3.7_a	2.0
Meinz and Salthouse (1998)	128	15.7	1.9
Salthouse (1996)	178	13.5	2.1
Salthouse (2001a), Study 1	222	13.6	2.3
Salthouse (2001a), Study 2	237	14.0	2.3
Salthouse (2001b)	206	16.0	2.0
Salthouse, Atkinson, and Berish (2003)	261	16.0	2.0
Salthouse and Ferrer-Caja (2003)	150	15.9	2.0
Salthouse, Fristoe, McGuthry, and Hambrick (1998), Study 2	191	15.0	2.3
Salthouse, Hambrick, Lukas, and Dell (1996), Study 2	77	15.6	2.1
Salthouse, Hancock, Meinz, and Hambrick (1996), Study 3	197	14.9	2.0
Salthouse, McGuthry, and Hambrick (1999)	189	15.4	2.3
Salthouse et al. (2000), Study 2	207	15.8	1.9
Salthouse, Toth, Hancock, and Woodard (1997)	124	15.3	2.2
Siedlecki, Salthouse, and Berish (2005)	330	15.7	2.0

Note. Education refers to the number of years of formal education completed except for those marked with a subscript *a*, which were classified as 1 (*less than 12 years*), 2 (*high school graduation*), 3 (*13–15 years of education*), 4 (*college graduate*), and 5 (*17 or more years of formal education*). Health is on a 5-point scale ranging from 1 (*excellent*) to 5 (*poor*).

Table 2
Description of Cognitive Tasks

Broad cognitive ability and test	*n*	Description	Source
Reasoning			
Matrix Reasoning	1,756	The participant selects the best alternative to complete the missing cell in a matrix.	Odd-numbered items from Raven (1962)
Cattell's Matrices	420	The participant selects the best of six alternatives to complete the missing cell of a 2 × 2 or 3 × 3 matrix.	Institute for Personality and Ability Testing (1973)
Figure Classification	459	Two or three groups of figures are presented at the top of the page, with figures within each group related in some way. Rows of figures are presented, and the participant marks which group the figures belong to.	Ekstrom, French, Harman, and Derman (1976)
Shipley Abstraction Test	420	The participant is given a series and responds with the number, letters, or word that completes the series.	Zachary (1986)
Letter Sets	796	The participant selects which groups of letters do not belong in each of 20 sets of letters.	Ekstrom et al. (1976)
Spatial visualization			
Spatial Relations	1,154	The participant mentally assembles an unfolded piece of paper and then determines which of four three-dimensional structures it most closely resembles.	Bennett, Seashore, and Wesman (1997)
Paper Folding	944	A piece of folded paper with a hole punched through the folded surface is presented, and the participant identifies the pattern of holes that would result when the paper is unfolded.	Ekstrom et al. (1976)
Form Boards	467	The participant selects the set of pieces that can be assembled to form a specified shape.	Ekstrom et al. (1976)
Memory			
WMS-III Free Recall	411	A list of 12 words is presented orally four times, with the participant recalling as many words as possible following each presentation. A second list of words is then presented and recalled, followed by an attempt to recall as many words as possible from the original list.	Wechsler (1997b)
Rey Auditory Learning Test	586	Fifteen words are read to the participant, followed immediately by a recall attempt. Five trials are given with the same list, with each trial consisting of a presentation and recall attempt.	Schmidt (1996)
Paired Associates	822	A set of six word pairs is presented orally. The participant receives a page containing the first member of each pair and responds with the second member. A second trial is then given with a new set of word pairs.	Salthouse, Fristoe, et al. (1998)
Speed			
Letter Comparison	3,182	The participant makes same-or-different judgments for pairs of letter strings as quickly as possible. Two pages of letter string pairs are presented, with 30 s allowed for each page.	Salthouse and Babcock (1991)
Pattern Comparison	3,182	Similar to Letter Comparison, except that instead of letter strings, the participants are presented with pairs of patterns composed of line segments.	Salthouse and Babcock (1991)
WAIS-III Digit–Symbol Substitution	411	Participants are allowed 2 min to write symbols below digits according to a code table displayed at the top of the page.	Wechsler (1997a)

Note. WMS-III = Wechsler Memory Scale–Third Edition; WAIS-III = Wechsler Adult Intelligence Scale–Third Edition.

tion, memory, and speed.[1] A short description of each task is given in Table 2. We describe the broad cognitive abilities in terms of Carroll's (1993) taxonomy: Reasoning consists of five inductive reasoning tasks; spatial visualization consists of three general visualization tasks; memory consists of two free recall tasks and one associative memory task; and speed consists of three perceptual speed tasks.

Vocabulary Tests

Four vocabulary tests, each with different formats, were used in this study. The Synonyms Vocabulary Test (abbreviated Synonyms Test; Salthouse, 1993) consists of 10 multiple-choice items, with five response alternatives for each item. Participants are instructed to circle the word whose meaning is most nearly the same as that of the target word, and the score is the total number of items answered correctly. The Antonyms Vocabulary Test (abbreviated Antonyms Test; Salthouse, 1993) is identical, except that participants are instructed to circle the word whose meaning is most nearly opposite to that of the target word. For both tests,

[1] The Wechsler Memory Scale–Third Edition Logical Memory (Wechsler, 1997b) also met selection criteria, but because it had different relations to age and other cognitive variables than the other memory variables, it was excluded from the analyses.

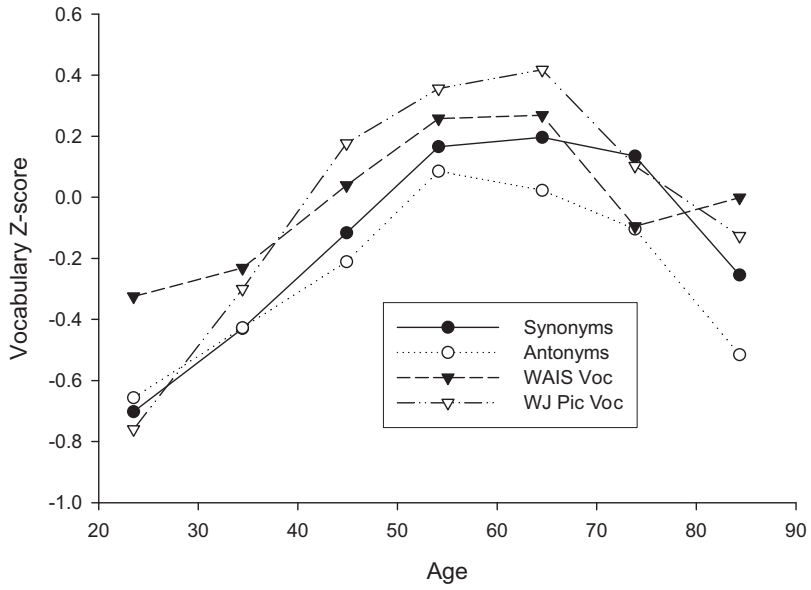

Figure 1. Relation between age and *z* scores on the vocabulary tests. WAIS Voc = Wechsler Adult Intelligence Scale–Third Edition Vocabulary test; WJ Pic Voc = Woodcock–Johnson Psycho-Educational Battery–Revised Picture Vocabulary test.

content was selected to be broadly representative of a number of sources, such as practice items on the Scholastic Aptitude Test, with no idiosyncratic selection mechanism.

The WAIS-III Vocabulary test (abbreviated WAIS Voc; Wechsler, 1997a) consists of 33 produce-the-definition items. Participants are given a target word and asked to define the word. Complete definitions are given an item score of 2, whereas incomplete definitions are given a partial credit item score of 1. Total score is the sum of the item scores. Scores on WAIS Voc were divided by 2 to maintain consistency with the other tests.

The WJ-R Picture Vocabulary test (abbreviated WJ Pic Voc) consists of 58 items. Participants are presented with a picture and respond with the name of the object depicted. To minimize testing time, only the final 30 items were administered. Responses are scored as either correct (score = 1) or incorrect (score = 0). Total score is the sum of the item scores.

Results

For all results described in this article, alpha was set to .01. We estimated all structural equation models with Amos (Arbuckle, 2006), using full information maximum likelihood estimation (Wothke, 2000), which allows for the analysis of incomplete data that is missing at random (R. J. A. Little & Rubin, 2002; McArdle, 1994). To maintain consistency in the vocabulary test scoring without affecting correlations among the variables, all vocabulary scores were converted to *z* scores based on the mean and standard deviation for the 739 persons who had complete data on all four tests. Fit of the four vocabulary tests to a single vocabulary factor was good (RMSEA = .037), and all standardized factor loadings were high (Synonyms Test = .94, Antonyms Test = .93, WAIS Voc = .87, WJ Pic Voc = .81).[2]

Age Relations

Age was positively related to scores on all four vocabulary tests, although the magnitude varied: all correlations equated, $\chi^2(3) = 124$, $p < .01$. The Synonyms Test ($r = .27$) and WJ Pic Voc ($r = .26$) were not significantly different in their correlation with age, $\chi^2(1) = 0.7$, $p = .40$. WJ Pic Voc and the Synonyms Test had stronger relations with age than the Antonyms Test ($r = .18$), $\Delta\chi^2(1) = 70$, $p < .01$, which in turn was more strongly related to age than WAIS Voc ($r = .09$): constrained equal to the Antonyms Test, $\Delta\chi^2(1) = 17$, $p < .01$. However, as shown in Figure 1, the age relations of all four tests were nonlinear and similarly shaped.

We fit a linear–linear spline model with fixed knot point to each of the vocabulary tests (Cudeck & Klebe, 2002). The linear–linear spline model consists of two linear trends, one before the knot point (linear growth) and one after (linear decline). We initially estimated the knot point separately for each vocabulary test, but to maximize comparability, we fixed the knot point at age 58, which was the approximate mean value and was within the 95% confidence interval for all four tests. We also considered a quadratic growth curve but opted for the linear–linear spline because it has easily interpreted parameters (growth and decline rates) that can be compared separately across the formats and because it fit at least marginally better than a quadratic model for all four tests and used the same number of parameters.

WJ Pic Voc had the strongest growth rate (.036 *SD*s per year before the knot point), followed by the Synonyms Test (.030), the

[2] A model with correlated residuals for the Synonyms Test and the Antonyms Test (i.e., multiple-choice method factor) yielded very similar results. A factor model with the formats residualized on the age functions (age before and after 58) described later was also not appreciably different.

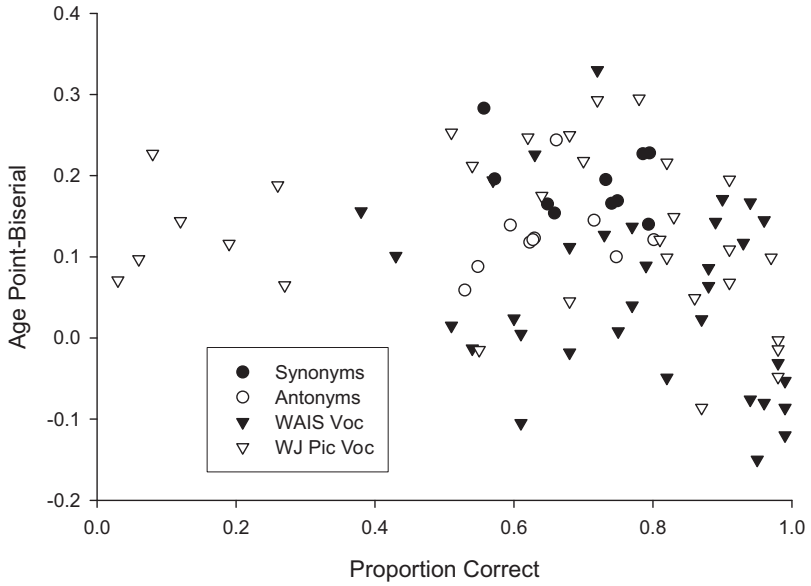

Figure 2. Relation between item difficulty and point-biserial correlations between age and items scores. Each point represents a single item. WAIS Voc = Wechsler Adult Intelligence Scale–Third Edition Vocabulary test; WJ Pic Voc = Woodcock–Johnson Psycho-Educational Battery–Revised Picture Vocabulary test.

Antonyms Test (.025) and WAIS Voc (.019). WJ Pic Voc also had the strongest decline rate (−.033), followed by WAIS Voc (−.025), the Antonyms Test (−.020), and the Synonyms Test (−.015). All age trends were different; equating the curves yielded a significant loss in fit compared to allowing all to be free, $\Delta\chi^2(6) = 147$, $p < .01$. Furthermore, equating any two curves yielded a significant loss in fit, smallest $\Delta\chi^2(2) = 18$, $p < .01$, for the Antonyms Test and WAIS Voc; largest $\Delta\chi^2(2) = 69$, $p < .01$, for the Synonyms Test and the Antonyms Test. These are not trivial differences; these results suggest that during the 40 years of adulthood before the knot point, the average age-related increase in WJ Pic Voc scores is 1.44 *SD,* whereas WAIS Voc is expected to increase only .76 SD, an effect size difference of .68 *SD.* At the other end of the life course, WJ Pic Voc is expected to decrease 1.32 *SD* over the approximately 40 years spanned by our study after the knot point, whereas the Synonyms Test is predicted to decrease only .60 *SD,* an effect size difference of .72 *SD.*

Artifactual Causes of Differential Age Relations

Three potentially artifactual sources of the differential age relations were examined before exploring relations to other cognitive variables. First, if scores on one test contained more measurement error than those on other tests, then the test scores would be less related to age even if the underlying latent trait has the same relation to age and to other cognitive abilities. However, we found no evidence of varying levels of measurement error; coefficient alpha was .85 for the Synonyms Test (n = 2,432), .86 for the Antonyms Test, .89 for the WAIS Voc (on the basis of these data; .93 as reported by Wechsler, 1997a), and .88 for the WJ Pic Voc (on the basis of these data; .88 as reported by McGrew, Werder, & Woodcock, 1991).[3]

A second possibility is that the tests vary in average difficulty and that they differ in their age relations because of difficulty variations rather than differences between the formats (Bowles, Grimm, & McArdle, 2005). If more difficult items, regardless of the type of test, are more negatively (or positively) related to age, then more difficult tests may be less (or more) strongly related to age. If this were the case, then the point-biserial correlation between age and item responses would be systematically related to item difficulty within each test. However, as displayed in Figure 2, there was consistency in the age point-biserial across item difficulty within tests. Furthermore, a regression of the point-biserial on item difficulty was not significant for any test.

A third possibility is order effects. Age relations may be, for example, stronger for tests presented later because of such irrelevant factors as age differences in rates of test fatigue. There was no consistent order of tests across studies with one exception: The Synonyms Test was administered immediately before the Antonyms Test. Thus, the age differences are unlikely to result from order effects.[4]

Relations to Other Cognitive Abilities

The baseline model we used is displayed in Figure 3. To assess relations of the vocabulary tests to other cognitive variables, we developed a factor analytic measurement model with five broad cognitive abilities—Vocabulary, Reasoning, Spatial Visualization, Speed, and Memory—and one higher order factor indicated by Reasoning, Spatial Visualization, Speed, and Memory and correlating with Vocabulary. Global fit statistics for this model were not

[3] Item-level data for the calculation of coefficient alpha was available only for approximately two thirds of the sample that took the Synonyms Test (n = 2,432) and the Antonyms Test (n = 2,408).

[4] No evidence was found of order effects within tests. Age point-biserials did not differ systematically within test for any format.

available because of the complex pattern of missing data, which did not allow for the estimation of a fully saturated model because some pairs of tests were never administered at the same time. However, all factor loadings on the broad cognitive abilities were positive and large, ranging from .74 to .90. The higher order factor was indistinguishable from Reasoning; the standardized factor loading was greater than one, and so we constrained the standardized factor loading to 1. For convenience, we named the higher order factor General Fluid Abilities (GFA), although we make no claim that it is identical to Horn and Cattell's Gf (Horn, 1985). Factor loadings for the other cognitive abilities on GFA were .89 for Spatial Visualization, .80 for Memory, and .79 for Speed; the correlation with Vocabulary was .44.

We then examined relations between the individual vocabulary tests and the broad cognitive abilities by adding paths from, in turn, GFA, the Spatial Visualization residual, the Memory residual, and the Speed residual to each of the four vocabulary tests. The paths from GFA tested whether the vocabulary tests had a different relation to GFA. The paths from the residuals tested whether the vocabulary tests had different relations to other cognitive abilities, independent of or controlling for GFA. GFA was differentially related to the four vocabulary formats. The standardized regression coefficients were .36 for the Synonyms Test, .44 for the Antonyms Test, .52 for WAIS Voc, and .39 for WJ Pic Voc.[5] All coefficients were significantly different except for the Synonyms Test and WJ Pic Voc, $\Delta\chi^2(1) = 1$, $p = .32$. The Spatial Visualization residual was more strongly related to WJ Pic Voc (.40) than to the other three vocabulary tests (.26), which were not significantly different from each other, $\Delta\chi^2(2) = 4$, $p = .14$. The Memory residual was more strongly related to WAIS Voc (.26) than to the other three vocabulary tests (.16), which were not significantly different from each other, $\Delta\chi^2(2) = 7$, $p = .03$. The Speed residual was more strongly related to the Antonyms Test (.19) than to the Synonyms Test or WAIS Voc (.09), which were not significantly different from each other, $\Delta\chi^2(1) < 1$, $p = .38$, and was least strongly related to WJ Pic Voc (−.06, which was not significantly different from 0, $p = .15$). These findings are summarized in Table 3 and illustrated in Figure 4.

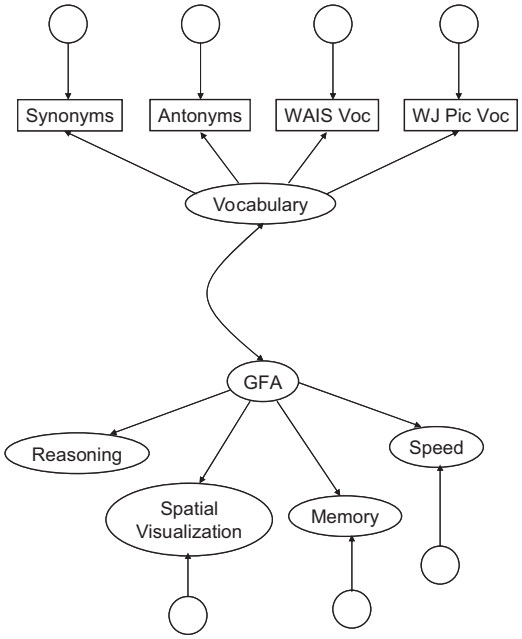

Figure 3. Structural equation measurement model. Indicators on the broad cognitive ability factors are suppressed. WAIS Voc = Wechsler Adult Intelligence Scale–Third Edition Vocabulary test; WJ Pic Voc = Woodcock–Johnson Psycho-Educational Battery–Revised Picture Vocabulary test; GFA = general fluid abilities.

Table 3
Relation of Vocabulary Tests to Broad Cognitive Abilities

Broad cognitive ability	Antonyms Test	WAIS Voc	WJ-R Pic Voc
GFA/reasoning	+	++	
Spatial visualization			+
Memory		+	
Speed	+		−

Note. Relations are relative to relations with the Synonyms Test. A plus sign indicates a stronger relation, a minus sign indicates a weaker relation, and two plus signs indicate a stronger relation than a single plus sign. WAIS = Wechsler Adult Intelligence Scale; Voc = Vocabulary; WJ-R = Woodcock–Johnson Psycho-Educational Battery–Revised; Pic = Picture; GFA = general fluid abilities.

We then developed a mediation model to examine whether the differential age relations can be accounted for by the differential relations to other cognitive variables. For example, does WAIS Voc have a different relation to age than the other three vocabulary tests because it is more strongly related to both GFA and Memory? If so, after including indirect effects of age through GFA and Memory, the direct effects of age should be identical to the direct effects of age for the other three vocabulary formats. We added to the measurement model in Figure 3 the two age variables (i.e., linear growth and linear decline with an age 58 knot point), with paths from the age variables to GFA, Spatial Visualization, Memory, Speed, and the four vocabulary tests. For statistical convenience and because it had the least complex pattern of relations to other broad cognitive variables, we set the Synonyms Test as our reference task, and as such, included paths from cognitive variables when the previous analyses indicated that the relations were significantly different from those for the Synonyms Test. These included paths from GFA to the Antonyms Test and WAIS Voc, from Spatial Visualization to WJ Pic Voc, from Memory to WAIS Voc, and from Speed to the Antonyms Test and WJ Pic Voc.

Results indicated that all age differences among the Synonyms Test, the Antonyms Test, and WAIS Voc were accounted for by differential relations to other cognitive variables. Constraining the path coefficients from the age variables to these three vocabulary tests did not yield significantly more misfit, $\Delta\chi^2(4) = 13$, $p = .02$. Confirming this result are the similar values for unconstrained direct (unmediated) effects of age (Synonyms Test = .030 growth and −.014 decline; Antonyms Test = .030 growth and −.008 decline; WAIS Voc = .025 growth and −.010 decline). WJ Pic Voc, on the other hand, still had a different age trend (.036 growth and −.032 decline), $\Delta\chi^2(2) = 27$, $p < .01$).

[5] For analyses addressing differential relations to GFA, the correlation between the Vocabulary factor and GFA was not included in the model.

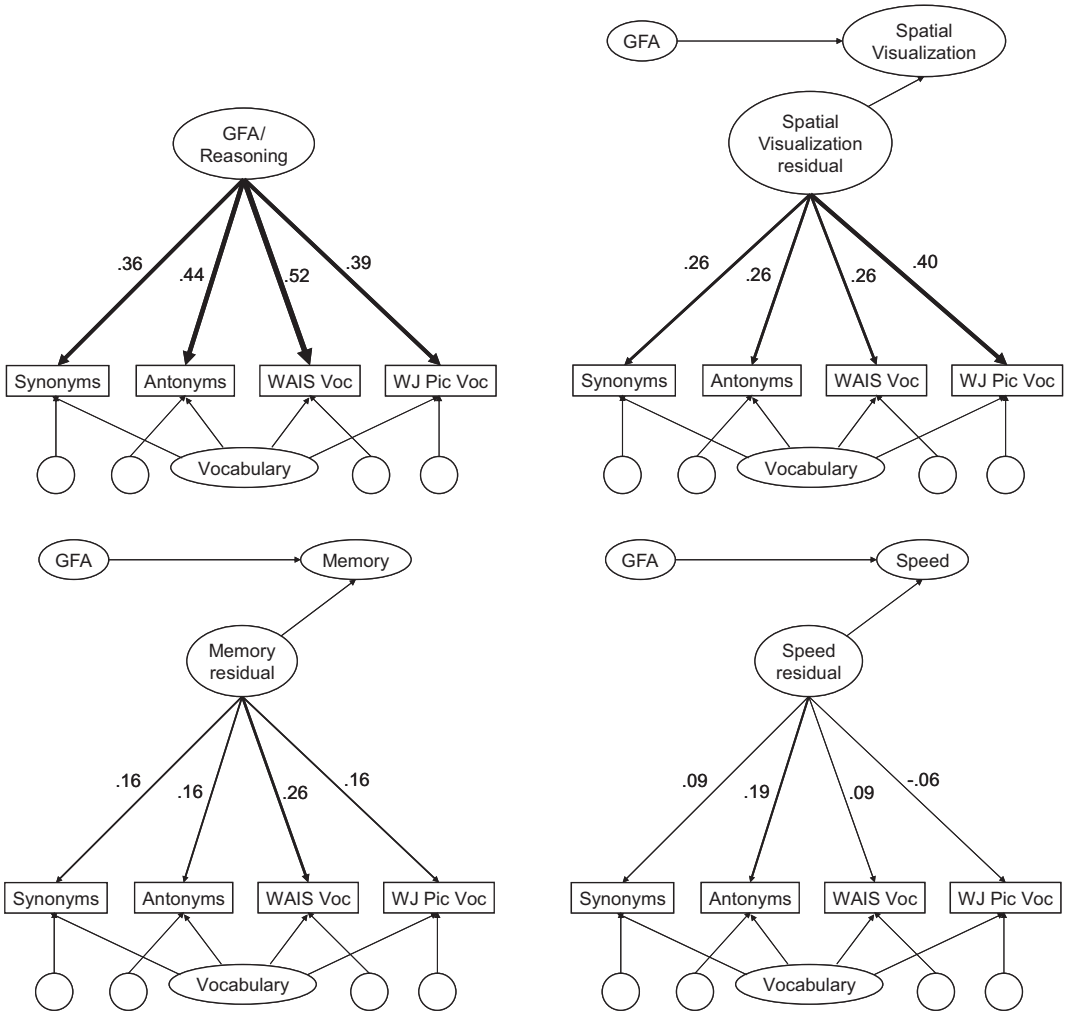

Figure 4. Abbreviated path models of the relation between cognitive abilities and vocabulary formats. Numbers and the thickness of the lines connecting to vocabulary tests represent standardized regression coefficients. GFA = general fluid abilities; WAIS Voc = Wechsler Adult Intelligence Scale–Third Edition Vocabulary test; WJ Pic Voc = Woodcock–Johnson Psycho-Educational Battery–Revised Picture Vocabulary test.

Discussion

Although different formats of vocabulary tests are generally considered interchangeable, they can have different relations to age and other cognitive abilities. Correlations with age differed substantially, ranging from .14 for the WAIS Voc produce-the-definition test to .30 for the WJ Pic Voc picture identification test. The age differences were still apparent when considering the age trends as nonlinear. In earlier adulthood (before the age 58 knot point), the WJ Pic Voc picture identification test had the strongest growth (.036 *SD*s per year, 1.44 *SD*s over 40 years), and the WAIS Voc produce-the-definition test had the weakest (0.019 *SD*s per year, 0.76 *SD*s over 40 years). In later adulthood (after age 58), WJ Pic Voc picture identification had the strongest decline (−0.033 *SD*s per year, −1.32 *SD*s over 40 years), and multiple-choice synonyms had the least (−0.015 *SD*s per year, −0.60 *SD*s over 40 years).

As summarized in Table 3, the tests had different relations to other cognitive variables. Compared to the multiple-choice Synonyms Test, WJ Pic Voc was more strongly related to spatial visualization and was negatively related to speed, the multiple-choice Antonyms Test was more strongly related to speed, and the WAIS Voc produce-the-definition test was more strongly related to reasoning and memory. For the most part, these relations to other broad cognitive abilities accounted for the differential age relations. However, WJ Pic Voc still had a stronger positive age-related growth and a stronger negative age-related decline than the other vocabulary tests.

Theories of the Aging of Vocabulary Knowledge

To account for these findings, a theory of the aging of vocabulary knowledge must include one or more cognitive processes that differ across format and are related to age. To our knowledge,

only two theories do this: the dual representation theory (McGinnis & Zelinski, 2000) and the spreading activation transmission deficit hypothesis (James & Burke, 2000; MacKay & Abrams, 1998; MacKay & Burke, 1990). Under dual representation theory, there are two cognitive representations of vocabulary knowledge, a detailed exact definition and a general gist, similar to the gist–verbatim distinction in memory (Brainerd & Reyna, 1992). Alternatively, there may be a continuum of specificity in multiple representations (McGinnis & Zelinski, 2003). Older adults are less able to generate and access the detailed definition (McGinnis & Zelinski, 2000) and compensate by relying more on the general representation (Botwinick & Storandt, 1974; Tun, Wingfield, Rosen, & Blanchard, 1998). Thus, there may be different age relations for different types of vocabulary tasks if the tasks differ in the sufficiency of the general representation for correct responses. Although this has been proposed as a theoretical explanation for differences in scores on the WAIS Voc (Botwinick & Storandt, 1974), there do not appear to be adequately detailed theoretical expectations of differences in the sufficiency of the general representation across tasks.

The spreading activation transmission deficit hypothesis (James & Burke, 2000; MacKay & Abrams, 1998; MacKay & Burke, 1990) suggests that the links between representations of a word and its definition or semantic meaning become weaker or less efficient with age (Burke, MacKay, & James, 2000; MacKay & Abrams, 1998). For production tasks, such as a produce-the-definition test, activation of the correct response (definition or target word) comes only from the stem of the vocabulary item (word or picture). In multiple-choice tasks, however, activation is passed not just from the target word, which weakens with age, but also from the multiple-choice options (Burke, MacKay, & James, 2000). Thus, the age-related degradation of the efficiency of the connections between nodes is more detrimental to production tasks than to multiple-choice tasks, a prediction confirmed in some research studies (Verhaeghen, 2003) and echoed in our results on later life declines on picture identification and produce-the-definition tasks than on multiple-choice antonym and synonym tasks.

When coupled with the WordNet theory (Gross, Fischer, & Miller, 1989; Gross & Miller, 1990), the transmission deficit hypothesis also predicts stronger declines for antonym knowledge than for synonym knowledge. According to WordNet theory, identifying an antonymous relationship between two words (e.g., *hot* and *cool*) involves identifying the direct or exact antonym of the first word (*hot* to *cold*) followed by recognizing a synonymous relationship between the second word and the direct antonym of the first (*cold* and *cool*). Thus, identifying antonyms requires traversing more links between nodes than identifying synonyms (Charles, Reed, & Derryberry, 1994; Gross et al., 1989), and therefore, under the transmission deficit hypothesis, antonym knowledge should be more susceptible to aging than synonym knowledge. This matches our finding that antonyms are more strongly negatively related to age in later adulthood.

Our findings suggest a third possible direction for theoretical development. The differences in age-related trends among multiple-choice synonyms, multiple-choice antonyms, and produce the definition (but not picture identification) were primarily accounted for by differential relations to reasoning or GFA. Reasoning declines after a peak age of approximately 20, so a vocabulary format more strongly related to reasoning would be expected to grow less during early adulthood and decline more in later adulthood. Our empirical results confirm this expectation: Produce the definition had weaker early adulthood growth and stronger later adulthood decline, and was most strongly related to reasoning, whereas multiple-choice synonym scores grew more rapidly during early adulthood, declined least in later adulthood, and were least strongly related to reasoning. Thus, differences among formats in the necessity or usefulness of reasoning may account for the differential age trends. However, theoretical explanations for why certain formats require or allow for more reasoning remain to be developed. It should be noted that reasoning as a theoretical explanation does not preclude the transmission deficit hypothesis or dual representation, which are at different levels of cognitive representation.

Processing Differences

Processing differences among different formats have been a key aspect of cognitive aging research. By developing task manipulations designed to tap different cognitive processes, researchers are able to isolate the hypothesized processing differences. For example, a great deal of research has been dedicated to understanding differences between recognition and recall in episodic memory (Craik & McDowd, 1987), leading to theories involving processes that are differentially related to age (e.g., Craik, 1983). Limited research has addressed such processing differences in vocabulary knowledge (e.g., Botwinick & Storandt, 1974; Burke, MacKay, & James, 2000; Verhaeghen, 2003). This study suggests some further directions that may be fruitful, by highlighting differences among test formats in their relations with broad cognitive abilities. These differences suggest that some of the formats require specific processes that are shared with other cognitive constructs. For example, the finding that antonyms knowledge shares more variance with speed than the other formats suggests that antonyms knowledge may require a process that is speed intensive. A systematic treatment of the processing differences is beyond the scope of this study, but we hope that it will provide an impetus for research addressing a more thorough understanding of the nature of vocabulary knowledge.

Interpretational Challenges

Although we interpret our findings as differences between formats, vocabulary test format is confounded by item content. Tests differed not only in format, but also in the particular target words. In addition, for the Synonyms Test and the Antonyms Test, the response options, and our findings, may be a result of idiosyncratic characteristics of the item content (Verhaeghen, 2003). It is not possible to remove this confound, because content necessarily varies with format. For example, the same response options could not be used for both the Synonyms Test and the Antonyms Test, even if the target words were identical. Item content differences are unlikely to be a major confound, however, because no item on any of the four tests was selected for content. Instead, the items were intended to be an essentially random selection from the large pool of potential vocabulary test items.

A second concern in the interpretation of our findings is the relation between Reasoning and the other broad cognitive factors.

The higher order factor, which we called GFA, was indistinguishable from the Reasoning factor, consistent with Gustafsson's (1984) assertion that Reasoning (or Gf) is identical to a General Abilities factor like Spearman's g. As a result, all variance shared among Reasoning, Spatial Visualization, Memory, and Speed was assigned to the GFA/Reasoning factor. This may have been caused by our selection of cognitive tasks, which consisted of a majority of reasoning tasks and very closely related spatial visualization tasks and, therefore, emphasized variance associated with reasoning. The differential relations of the vocabulary tests to other cognitive variables hold regardless of the exact identity of the higher order factor, although the interpretation of the differential relations to GFA/Reasoning remains somewhat uncertain.

A third concern is that we used two locally developed vocabulary tests: the multiple-choice Synonyms Test and the multiple-choice Antonyms Test. These tests may have distinctive characteristics that limit the generalizability of our results. The multiple-choice Synonyms Test was identical in format to such well-used multiple-choice vocabulary tests as the Shipley Vocabulary test (Shipley, 1946) or the Thorndike–Gallup test of verbal knowledge (Thorndike & Gallup, 1944). To our knowledge, no standard cognitive test battery contains a test of antonyms knowledge; the format, however, was identical to the Synonyms Test except for the request for a word opposite in meaning instead of identical in meaning. Content of the items on both tests was adapted from a number of sources with no idiosyncratic selection mechanism. Furthermore, dividing the tests into two subtests (odds and evens) and analyzing each subtest separately yielded the same statistical results. Therefore, we do not expect that our results are specific to these particular multiple-choice tests.

A final interpretational concern is that we used standard scoring procedures for all test formats and considered effects only at the test score level. There is evidence that, at least on one vocabulary test not included in this study, there may be differential relations with age depending on the particular items considered (Bowles, Grimm, & McArdle, 2005). We did not find systematic differences across items in age point-biserials within any tests, suggesting that the test score was an appropriate level at which to consider differential age relations. Nonetheless, we consider this a topic for future research into the generalizability of both our findings and those of Bowles et al. (2005).

Measurement Issues

No instrument is either an exclusive or an exhaustive representation of a latent construct. Rather, there are many ways to measure a construct, and no particular way can completely reflect the construct of interest (Hand, 2004). Recognizing this conceptual distinction highlights two critical aspects of measurement. First, it is important to identify how each instrument measures the construct and the manner in which each instrument measures the construct differently from other instruments measuring the same construct. Each instrument may require different cognitive processes, and identification of these processes may inform understanding of the construct. These processes can be identified and understood only in the context of other variables, through evidence of convergence and divergence with other instruments measuring the same construct, and through evidence of convergence and divergence with other constructs. The current study provides an example of this process.

Second, as this study highlights, it is important to have multiple indicators of the same construct in order to assess the breadth of the construct. The call for multiple indicators is certainly not new (T. Little, Lindenberger, & Nesselroade, 1999), but these results illustrate one reason why having only a single indicator could lead to misleading results. A single indicator may involve processes that are not involved in other instruments measuring the same construct. Findings about relations with other variables or constructs could therefore reflect those processes unique to the single indicator instead of the common processes that define the construct. For example, using the WJ Pic Voc test as the only indicator of vocabulary knowledge may overestimate the age-related growth and decline of vocabulary knowledge, whereas using only the WAIS Voc test may overestimate the relation between vocabulary knowledge and reasoning.

Practical Implications

Despite the importance of using multiple indicators of vocabulary knowledge, it is often impractical to use more than one. We argue that, although there can never be a pure measure of vocabulary knowledge, a multiple-choice test of synonyms offers the closest approximation. The Synonyms Test had the highest factor loading on a Vocabulary factor (.94). The Synonyms Test was also the least different from the other tests; it had the fewest significant differences from the other vocabulary tests in terms of relations to other cognitive abilities. Therefore, it was closest to an average test. We also suggest that WJ Pic Voc may not be a good choice for a single vocabulary indicator, because of its lowest factor loading (.81), its differential relations to other cognitive abilities, and the unclear source of its differential age relations.

Summary

Despite Carroll's (1993) claim of *indifference of the indicator* for vocabulary tests, different formats can have different relations to age and to other cognitive abilities. Of the four formats we examined, multiple-choice synonyms, multiple-choice antonyms, and produce the definition tests had the same age trends after accounting for differential relations to other cognitive variables, primarily reasoning. A theoretical explanation for the differences may therefore include differences in the role of reasoning. Picture identification, however, has a different age trend that is not accounted for by other variables, and may not be a good choice of format when a single indicator of vocabulary is used. Until greater understanding of vocabulary knowledge is gained, it is strongly suggested that researchers include multiple indicators of vocabulary knowledge, especially when vocabulary knowledge is an important focus of the research.

References

*References marked with an asterisk indicate studies included in the data.

Alwin, D. F., & McCammon, R. J. (2001). Aging, cohorts, and verbal ability. *Journal of Gerontology: Social Sciences, 56*(B), S151–S161.

Arbuckle, J. L. (2006). Amos (Version 7) [Computer software]. Spring House, PA: Amos Development Corporation.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1997). *Differential Aptitude Test.* San Antonio, TX: The Psychological Corporation.

Botwinick, J., & Storandt, M. (1974). Vocabulary ability in later life. *Journal of Genetic Psychology, 125,* 303–308.

Bowles, R. P., Grimm, K. J., & McArdle, J. J. (2005). A structural factor analysis of vocabulary knowledge and relations to age. *Journal of Gerontology: Psychological Sciences, 60,* P234–P241.

Brainerd, C. J., & Reyna, V. F. (1992). Explaining "memory free" reasoning. *Psychological Science, 3,* 332–339.

Burke, D. M., MacKay, D. G., & James, L. E. (2000). Theoretical approaches to language and aging. In T. J. Perfect & E. A. Maylor (Eds.), *Models of cognitive aging* (pp. 204–237). Oxford, England: Oxford University Press.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge, England: Cambridge University Press.

Charles, W. G., Reed, M. A., & Derryberry, D. (1994). Conceptual and associative processing in antonymy and synonymy. *Applied Psycholinguistics, 15,* 329–354.

Craik, F. I. M. (1983). On the transfer of information from temporary to permanent memory. *Philosophical Transactions of the Royal Society: B. Biological Sciences, 302,* 341–359.

Craik, F. I. M., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 474–479.

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. *Psychological Methods, 7,* 41–63.

Ekstrom, R. B., French, J. W., Harman, H. H., & Derman, D. (1976). *Kit of factor-referenced cognitive tests.* Princeton, NJ: Educational Testing Service.

Gross, D., Fischer, U., & Miller, G. (1989). Antonymy and the representation of adjectival meanings. *Memory and Language, 28,* 93–106.

Gross, D., & Miller, K. J. (1990). Adjectives in WordNet. *International Journal of Lexicography, 3,* 265–277.

Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8,* 179–203.

∗Hambrick, D. Z., Salthouse, T. A., & Meinz, E. J. (1999). Predictors of crossword puzzle proficiency and moderators of age–cognition relations. *Journal of Experimental Psychology: General, 128,* 131–164.

Hand, D. J. (2004). *Measurement theory and practice: The world through quantification.* London: Arnold.

Horn, J. L. (1985). Remodeling old models of intelligence: Gf-Gc theory. In B. B. Wolman (Ed.), *Handbook of intelligence* (pp. 267–300). New York: Wiley.

Institute for Personality and Ability Testing. (1973). *Measuring intelligence with the culture fair tests.* Champaign, IL: Author.

James, L. E., & Burke, D. M. (2000). Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1378–1391.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Little, T., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables. *Psychological Methods, 4,* 192–211.

MacKay, D. G., & Abrams, L. (1998). Age-linked declines in retrieving orthographic knowledge: Empirical, practical, and theoretical implications. *Psychology and Aging, 13,* 647–662.

MacKay, D. G., & Burke, D. M. (1990). Cognition and aging: A theory of new learning and the use of old connections. In T. M. Hess (Ed.), *Aging and cognition: Knowledge organization and utilization* (pp. 213–263). New York: Elsevier.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29,* 409–454.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2008). *Modeling latent growth curves using longitudinal data with non-repeated measurements.* Manuscript submitted for publication.

McGinnis, D., & Zelinski, E. M. (2000). Understanding unfamiliar words: The influence of processing resources, vocabulary knowledge, and age. *Psychology and Aging, 15,* 235–250.

McGinnis, D., & Zelinski, E. M. (2003). Understanding unfamiliar words in young, young–old, and old–old adults: Inferential processing and abstraction deficits. *Psychology and Aging, 18,* 497–509.

McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). *WJ-R technical manual.* Allen, TX: DLM.

McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual, Woodcock–Johnson III.* Itasca, IL: Riverside.

∗Meinz, E. J., & Salthouse, T. A. (1998). The effects of age and experience on memory for visually presented music. *Journal of Gerontology: Psychological Sciences, 53*(B), P60–P69.

Munoz-Sandoval, A. F., Cummins, J., Alvarado, C. G., & Ruef, M. L. (1998). *Bilingual verbal abilities test: Comprehensive manual.* Itasca, IL: Riverside.

Raven, J. (1962). *Advanced Progressive Matrices.* London: H. K. Lewis.

Salthouse, T. A. (1993). Speed and knowledge as determinants of adult age differences in verbal tasks. *Journal of Gerontology: Psychological Sciences, 48,* P29–P36.

∗Salthouse, T. A. (1996). General and specific speed mediation of adult age differences in memory. *Journal of Gerontology: Psychological Sciences, 51*(B), P30–P42.

∗Salthouse, T. A. (2001a). Attempted decomposition of age-related influences on two tests of reasoning. *Psychology and Aging, 16,* 251–263.

∗Salthouse, T. A. (2001b). Structural models of the relations between age and measures of cognitive functioning. *Intelligence, 29,* 93–115.

∗Salthouse, T. A., Atkinson, T. M., & Berish, D. E. (2003). Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of Experimental Psychology: General, 132,* 566–594.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology, 27,* 763–776.

∗Salthouse, T. A., & Ferrer-Caja, E. (2003). What needs to be explained to account for age-related effects on multiple cognitive variables. *Psychology and Aging, 18,* 91–110.

∗Salthouse, T. A., Fristoe, N., McGuthry, K. E., & Hambrick, D. Z. (1998). Relation of task switching to speed, age, and fluid intelligence. *Psychology and Aging, 13,* 445–461.

∗Salthouse, T. A., Hambrick, D. Z., Lukas, K. E., & Dell, T. C. (1996). Determinants of adult age differences on synthetic work performance. *Journal of Experimental Psychology: Applied, 2,* 305–329.

∗Salthouse, T. A., Hancock, H. E., Meinz, E. J., & Hambrick, D. Z. (1996). Interrelations of age, visual acuity, and cognitive functioning. *Journal of Gerontology: Psychological Sciences, 51*(B), P317–P330.

∗Salthouse, T. A., McGuthry, K. E., & Hambrick, D. Z. (1999). A framework for analyzing and interpreting differential aging patterns: Application to three measures of implicit learning. *Aging, Neuropsychology, and Cognition, 6,* 1–18.

∗Salthouse, T. A., Toth, J., Daniels, K., Parks, C., Pak, R., Wolbrette, M., & Hocking, K. J. (2000). Effects of aging on efficiency of task switching in a variant of the trail making test. *Neuropsychology, 14,* 102–111.

∗Salthouse, T. A., Toth, J., Hancock, H. E., & Woodard, J. L. (1997). Controlled and automatic forms of memory and attention: Process purity and the uniqueness of age-related influences. *Journal of Gerontology: Psychological Sciences, 52*(B), P216–P228.

Schaie, K. W. (1996). *Intellectual development in adulthood: The Seattle longitudinal study.* Cambridge, England: Cambridge University Press.

Schmidt, M. (1996). *Rey Auditory Verbal Learning Test: A handbook.* Los Angeles: Western Psychological Services.

Shipley, W. C. (1946). *Institute of Living Scale.* Los Angeles: Western Psychological Services.

∗Siedlecki, K. L., Salthouse, T. A., & Berish, D. E. (2005). Is there anything special about the aging of source memory? *Psychology and Aging, 20,* 19–32.

Singer, T., Verhaeghen, P., Ghisletta, P., Lindenberger, U., & Baltes, P. B. (2003). The fate of cognition in very old age: Six-year longitudinal findings in the Berlin Aging Study (BASE). *Psychology and Aging, 18,* 318–331.

Sorenson, H. (1938). *Adult abilities.* Minneapolis, MN: University of Minnesota Press.

Spearman, C. (1927). *The abilities of man: Their nature and measurement.* New York: Macmillan.

Thorndike, R. L., & Gallup, G. H. (1944). Verbal intelligence in the American adult. *Journal of General Psychology, 30,* 75–85.

Tun, P. A., Wingfield, A., Rosen, M. J., & Blanchard, L. (1998). Response latencies for false memories: Gist-based processes in normal aging. *Psychology and Aging, 13,* 230–241.

Verhaeghen, P. (2003). Aging and vocabulary score: A meta-analysis. *Psychology and Aging, 18,* 332–339.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale–Third Edition (WAIS-III).* San Antonio: Harcourt Assessment.

Wechsler, D. (1997b). *Wechsler Memory Scale–Third Edition (WMS-III).* San Antonio, TX: Harcourt Assessment.

Woodcock, R. W. (1987). *Woodcock Reading Mastery Test.* Circle Pines, MN: American Guidance Service.

Woodcock, R. W., & Johnson, M. B. (1990). *Woodcock–Johnson Psycho-Educational Battery–Revised.* Chicago, IL: Riverside.

Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data* (pp. 219–240). Mahwah, NJ: Erlbaum.

Zachary, R. A. (1986). *Shipley Institute of Living Scale: Revised manual.* Los Angeles: Western Psychological Services.

Received November 29, 2007
Revision received March 28, 2008
Accepted April 9, 2008 ■